“Thinking Statistically”: probability& statistics taught with an emphasis on using simulations and re-sampling methods to both DO statistics (analyze data) and understand concepts.

Three motivations:

1. Most existing statistics courses are simultaneously too superficial (e.g. flow-chart approach to test selection) and too detailed (e.g. calculations of covariance), and fail to develop intuitions that are central to statistical thinking (e.g. What is a p-value? a Bayes Factor?).
2. There is an abundance of good web-based material that can be taken advantage of if the students understand basic concepts and have basic programming skills.
3. We are not teaching our students the statistical approaches they will actually need in the world of big data. c.f. Efron, “Large-scale inference.”

Quarter course: eight 2-hour class sessions

Approach: Prior to class, students will view 2-4 hours of lectures on-line from Dr. Brian Healy’s biostatistics course. In class, we will focus on demos to build intuition and practice different approaches to analyzing real data sets, with a strong emphasis on resampling methods (i.e. bootstrap and permutation tests). All in-class programming will be done in MATLAB. Students must have either taken Neurobiology 306qc: Quantitative Methods for Biologists or demonstrate proficiency to the instructor.

**Syllabus**

Week #1: Data and Probability

Video: 1: Introduction to Biostatistics, 2: Basics of Probability

In class: distribution plots vs. box plots; PIN data visualization; dual-code for image analysis

In class: Bayes vs. Frequentist; Bayes dice demo; central limit theorem demo; James-Stein demo

Reading: Leek, “How to share data with a statistician”; Wickham, “Tidy Data”; Allen et al., Neuron 2012

Intro (15):

* Why this course?
* Importance of becoming a life-long statistician: e.g. subscribe to datacolada.org
* Course mechanics and expectations

Data (45)

* How to share data with a statistician. tidy data / clean data
* View your data in the rawest form that is visually comprehensible: PIN data
* Think hard about effective ways to visualize your data: PIN data cont’d
* In publications: Hide less; show more (Allen et al. 2012):
  + distribution plots (a.k.a. ‘violin’ plots) vs. whisker plots vs. bar plots
  + matrix of covariate plots
  + dual-code colorbars for images
* Sampling error vs. bias
  + distorted ruler class exercise
  + XLM example: meng-statistics\_perils\_of\_self-selection.pdf (1 page)

Probability (60)

* using simulations to solve (difficult!) probability problems
  + class picks number ‘randomly’ [1,4]
  + extension to birthday problem
* using simulations to gain intuition
  + CLT demo
* Frequentist vs. Bayes from 40k feet
  + Bayes theorem: 4 views
    - rule for inverting conditional probabilities
    - method for combining new data with existing knowledge
      * twin problem: U/S shows 2 boys; 1/3 twins identical; P(twins/US)?
    - method for updating beliefs as evidence accumulates over time
      * Bayes dice demo (Bayes factors, etc.)
    - a religion (compare and contrast w/ frequentist)
  + Converting p-values to Bayes factors? [prob. too early & not enough time]

Week #2: Bootstrap I: Hypothesis testing, Standard error and confidence intervals

Video: [Could we do basics of probability week #1, then do intro to biostats week #2?]

In class: introduction to bootstrap; bootstrapping standard errors and confidence intervals; ‘etMouseCI.m’, ‘etMouse2SampleStats.m’,

Week #3: Bootstrap II:

In class: Resampling for hypothesis tests and CI hypothesis testing: permutation test (‘etMouse2SamplePermutationTest.m’) vs. bootstrap test (‘etMouse2SampleBootstrapTest.m’)

Week #4: Power, P-hacking and Reproducibility (“torturing data ethically” – XLM)

;

In class: multiple comparisons; “researcher degrees of freedom”; simulation of stopping rule for data collection (‘dfSim2.m’); power calculations via simulation (‘powSimTtest2.m’)

Week #5: Linear Regression I

Video: 7. Linear regression and correlation; 12. Regression diagnostics

In class: basics of MATLAB regression tools: regress, glmfit, fitglm; ‘etHormoneRegression.m’, ‘etCellSurvivalReg.m’ (bootstrapping SEs when using other than LSE to do regression)

Week #6: Linear Regression II

Video: 11. Multiple Linear Regression I; 16/17. Logistic Regression

In class: GLM logistic regression on microstim experiment data (‘mStimLogisticRegressionDemo.m’)

Week #7: Linear Regression III

In class: GLM to fit Poisson point process model to place cell data (‘placeCellFitEx.m’)

In class: over-fitting (‘OverFit.m’); cross-validation ; regularization

Week #8: Introduction to Machine Learning: dimensionality reduction, classifiers

Video: Jan?; Chris Harvey NB204 lecture?

Video: https://www.youtube.com/watch?v=\_UVHneBUBW0

In class: spike sorting; Novembre et al. 2008?; Stephen Holtz & Alex Batchelor exercise: PCA neurons

In class: gradient descent using an objective function; regression as a classifier; SVM

Week #8alt: Intro to Empirical Bayes: False Discovery Rates

Video: 33. Multiple comparisons

In class: ‘MultipleComparisonsExercise.m’; Stein’s paradox (‘JSdemo.m’); distribution of p-values under H0; P-curve; ‘FDRdemo.m’

Week #8altalt: Experimental design: every decision one makes about how to do the experiment has a major effect on how to choose the appropriate analysis

General Resources:

Datacolada: <http://datacolada.org/>

[MIT 18.05](https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/index.htm)

Nature’s statistics for biologists: <http://www.nature.com/collections/qghhqm/content/practical-guides>

BH notes:

big picture idea: entire video block prior to course: 1-7, selected regression

1 month lead time for packaging of “desire2learn” back-bone (ask Tari about HMS requirements and canvas vs. desire2learn)